* Q12.2 – Give an application example of where the border between normal objects and outliers is often unclear, so that the degree to which an object is an outlier has to be well estimated.

There are a lot of medical and clinical applications that often result in unclear borders between normal and outlier objects. For example, in cancer detection / prevention, it is critical to monitor cell growth rates among individuals. Therefore, it is vital to be able to get a baseline value for an individual and/or group of individuals and also capable of estimating an outlier among very small deviations among individuals. This can be difficult, because individuals are all different and have slight variations in cell growth rates and compositions. This example describes cancer detection; however, it can be applied to a wide-range of medical conditions: degenerative eye conditions, spread of diseases, genetic anomalies, etc. It is critical in all these applications to be able to estimate outliers with precision.

* Q12.3 – Adapt a simple semi-supervised method for outlier detection. Discuss the scenario where you have (a) only some labeled examples of normal objects, and (b) only some labeled examples of outliers.

1. If the labels are only provided for some examples of normal objects, then a model can be used on the identified objects to label nearby unidentified objects as normal. These models can be statistical, proximity-based, or clustering-based methods. Objects found outside of these clusters/groups can then be identified as outliers based on the method used.
2. If the labels are only provided for some examples of outlier objects, then it will be more difficult create an effective model solely based on a semi-supervised model. This is the case, because it is unlikely that the labeled outliers represent all types of outliers present in the data set. Therefore, it is more effective to use a combination of semi-supervised to identify similar outliers to the ones provided; however, unsupervised methods will also need to be used to identify the clusters of normal objects. In conjunction, these methods can be effective at creating effective models when only a few outlier labels are provided.

* Q12.4 – Using an equal-depth histogram, design a way to assign an object an outlier score.

In an equal-depth histogram, we can utilize the range of values in each bin to create an object outlier score. For the example, consider the following histogram bins:

* + Bin 1 (Values Range: 0 – 50) (Frequency 10)
  + Bin 2 (Values Range: 51 – 60) (Frequency 10)
  + Bin 3 (Values Range: 61 – 65) (Frequency 10)
  + Bin 4 (Values Range: 66 – 100) (Frequency 10)

We can use the ratio of the range of values covered by a bin to the total range of values covered by the histogram to create an outlier score for a value. For example, if we are given a value of 8, this would fall into Bin 1 (which has a range that covers 50% of all the values). Therefore, 1/0.50 = 2. Another value, 62 falls in Bin 3 (which has a range that covers 4% of all the values). Therefore, 1/0.04 = 25. In this system, the lower the value the more likely it is an outlier. As a result, a value of 8 would be considered more of an outlier than 62 in this situation.

* Q12.5 – Consider the nested loop approach to mining distance-based outliers (Figure 12.6). Suppose the objects in a data set are arranged randomly, that is, each object has the same probability to appear in a position. Show that when the number of outlier objects is small with respect to the total number of objects in the whole data set, the expected number of distance calculations is linear to the number of objects.

In most cases, the inner loop of the nest loop algorithm can be terminated early, because π \* n objects were found within a distance r from oi. This relationship is linear as the total number of objects increases, because the ratio value selected should stay the same. Therefore, as the number of objects increase (e.g. from 10 to 100 to 1,000), then the loop must find the same ratio of objects to meet the criteria of the loop before terminating and concluding it is not an outlier. If the ratio is set at 0.50, then the inner loop would terminate early once it calculates the distance for 5 objects, 50 objects, and 500 objects, respectively. As a result, this illustrates the linear relationship between the nested loop approach and increasing the number of objects in the data set.

* Q12.6 – In the density-based outlier detection method of Section 12.4.3, the definition of local reachability density has a potential problem: lrdk(o) = ∝ may occur. Explain why this may occur and propose a fix to the issue.

The value for local reachability density could approach infinity when the distance between objects in a neighborhood is extremely small (approaching 0). Therefore, when you divide by the reachdistk, you will begin to approach infinity. As a result, if a point is located in the exact same spot, then you will return a value of infinity. A proposed fix to this issue would be to add a new variable/parameter that must be greater than 0 to eliminate the possibility of dividing by a value of 0.

